UNINPAHU

MACHINE-LEARNING-I

Autores:

Angie López

Marlon Peña

Marlon Albarracín

Bogotá D.C.

25 de junio del 2025

Índice

Capítulo 1

1. Análisis Profundo de Regresión Lineal Simple: Horas de Trabajo vs. Unidades Producidas

2. Introducción

En el presente estudio se analiza la relación entre el tiempo dedicado al trabajo (en horas) y la cantidad de unidades producidas. La hipótesis subyacente es que a mayor número de horas trabajadas, mayor será la producción. Este tipo de análisis es fundamental para áreas como administración de operaciones, productividad laboral y gestión del rendimiento en fábricas, talleres u oficinas. Para ello, se emplea un modelo de regresión lineal simple, que permite cuantificar y predecir el comportamiento de la variable dependiente (producción) a partir de la variable independiente (horas trabajadas).

El análisis se ha realizado utilizando tres enfoques diferentes para estimar el modelo:

- Gradiente descendente (método iterativo de optimización).
- `scikit-learn` (biblioteca estándar de machine learning en Python).
- `statsmodels` (enfoque estadístico clásico con pruebas de hipótesis).

Además, se validan los supuestos clásicos del modelo lineal (normalidad, homocedasticidad e independencia de los errores) para asegurar la validez del modelo desde el punto de vista inferencial.

3. Exploración de los Datos

El conjunto de datos contiene dos variables numéricas:

- **Trabajo (x)**: horas de dedicación laboral.
- **Producción (y)**: número de unidades fabricadas en ese tiempo.

El primer paso fue realizar una inspección visual a través de un diagrama de dispersión, el cual mostró una relación positiva aproximadamente lineal. Esto sugiere que puede ser apropiado aplicar un modelo lineal simple.

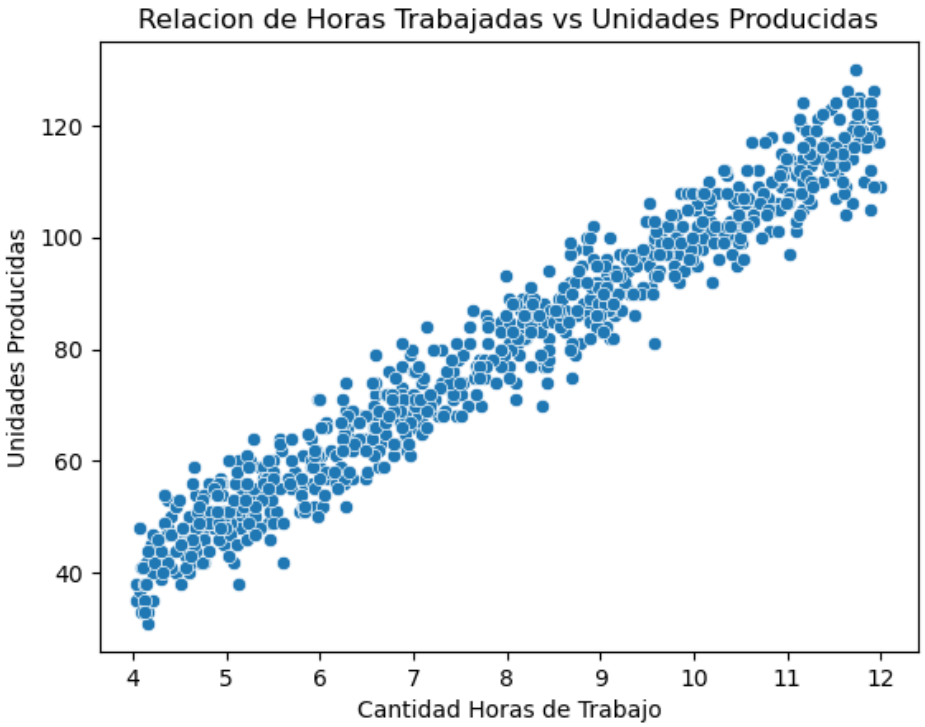


Figura 1: Relación entre Horas de Trabajo y Producción

Regresión Lineal Simple.

\n 1. Gráfico de dispersión – Relación de Horas Trabajadas vs Unidades Producidas

Qué muestra: Una nube de puntos que indica cómo se relacionan las horas trabajadas con la cantidad de unidades producidas.

Interpretación:

- Hay una relación lineal positiva clara: a mayor cantidad de horas trabajadas, mayor producción.
- Los puntos están bastante agrupados alrededor de una línea imaginaria, lo que sugiere un fuerte ajuste lineal.

4. Modelo de Regresión Lineal Simple

Se busca ajustar un modelo de la forma:

$$y = \theta_0 + \theta_1 x + \varepsilon$$

donde:

- y: unidades producidas (variable dependiente).
- x: horas trabajadas (variable independiente).
- θ_0: intercepto (producción cuando el tiempo de trabajo es cero).
- θ_1: pendiente (incremento de la producción por cada hora extra trabajada).
- ε: término de error aleatorio.

4.1. Estimación por Gradiente Descendente

El gradiente descendente es un algoritmo de optimización numérica que permite encontrar los valores óptimos de los parámetros del modelo al minimizar la función de coste, típicamente el error cuadrático medio (ECM).

Tras múltiples iteraciones, los valores convergieron a:

$$\hat{\theta}_0 = 3{,}9177, \quad \hat{\theta}_1 = 17{,}142$$

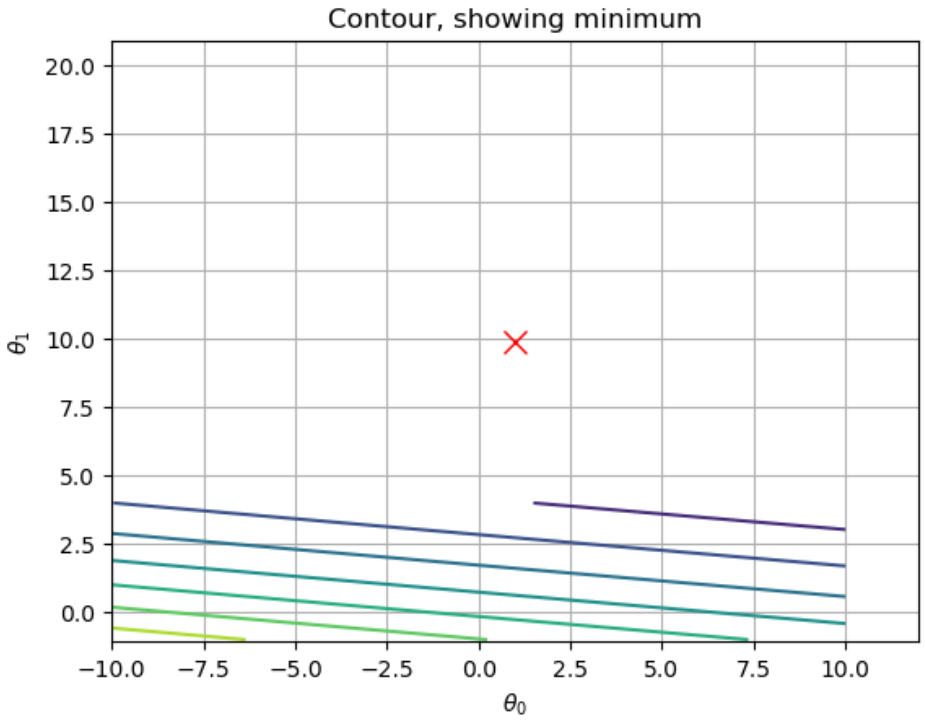


Figura 2: Curvas de nivel de la función de coste en gradiente descendente

\n 2. Gráfico de contorno – Mínimo de la función de costo (Gradiente Descendente)

Qué muestra: Este gráfico representa las curvas de nivel (contour) de la función de costo utilizada para encontrar los mejores parámetros de la regresión (,).

Interpretación:

- El eje X representa (intercepto), y el eje Y (pendiente).
- El punto rojo marca el mínimo de la función de costo: el punto óptimo donde el error es menor.
- Esto confirma que el gradiente descendente funcionó correctamente y encontró los mejores parámetros del modelo.

4.2. Estimación con `scikit-learn`

El modelo ajustado mediante la función `LinearRegression()` de `scikit-learn` arrojó los mismos coeficientes, validando el resultado anterior desde otra perspectiva:

$$\hat{\theta}_0 = 3{,}9177, \quad \hat{\theta}_1 = 17{,}142$$

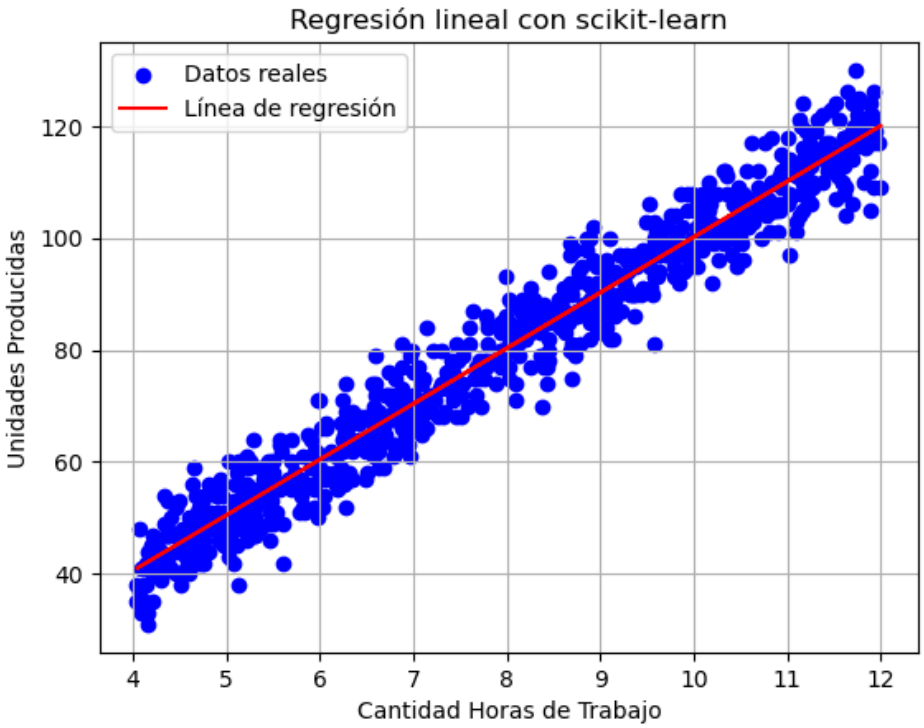


Figura 3: Modelo de regresión ajustado con `scikit-learn`

\n 3. Regresión lineal con Scikit-learn

Qué muestra:

- Los puntos azules representan los datos reales.
- La línea roja es la recta de regresión generada automáticamente con la librería Scikit-learn.

Resultados: Intercepto y pendiente.

Interpretación:

- Se confirma la relación positiva: por cada hora adicional de trabajo, se producen aproximadamente 17.14 unidades más.
- La línea ajusta muy bien a los datos → buen modelo predictivo.

4.3. Estimación con `statsmodels`

El uso de `statsmodels` permite realizar una estimación estadística con pruebas de hipótesis e intervalos de confianza. A continuación, se presentan los resultados:

Parámetro	**Coeficiente**	**Error Est.**	***p*-valor**	**IC 95 %**
Intercepto (θ_0)	3.9177	0.460	¡0.001	[3.01, 4.82]
Pendiente (θ_1)	17.142	0.992	¡0.001	[16.81, 17.47]

Cuadro 1: Estimaciones obtenidas con `statsmodels`

$$R^2 = 0{,}956$$

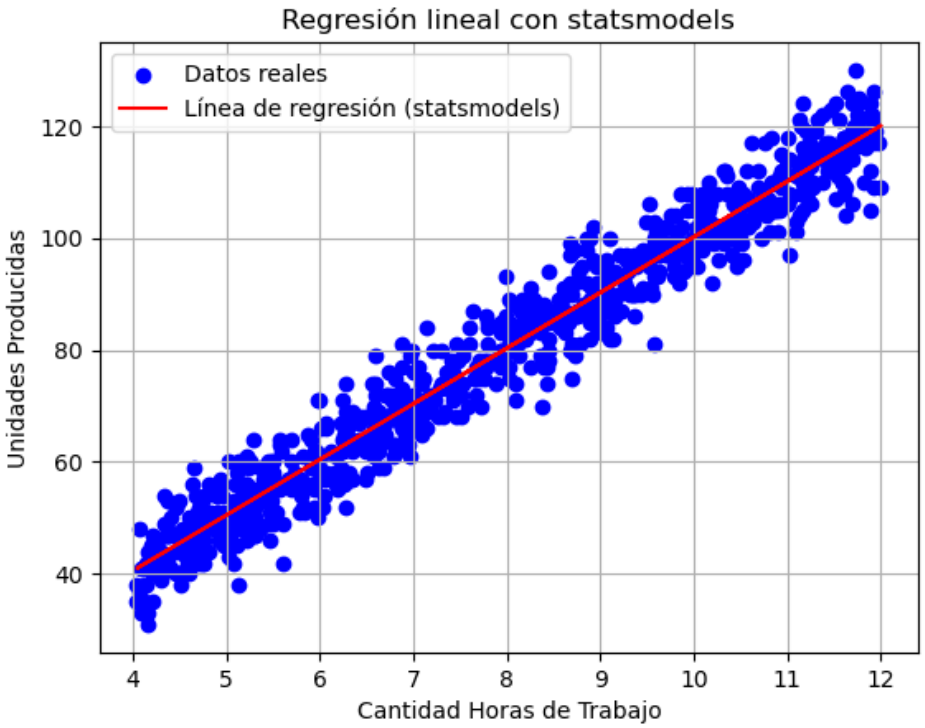


Figura 4: Línea de regresión ajustada con `statsmodels`

\n 4. Regresión lineal con Statsmodels

Qué muestra:

- Igual que el anterior, pero usando la librería Statsmodels, que permite obtener estadísticas más detalladas del modelo.

Resultados:

- Misma recta de regresión (coeficientes idénticos).
- Además del gráfico, Statsmodels ofrece información clave: p-valores, intervalos de confianza, R^2, etc.

Interpretación:

- $R^2 = 0{,}956 \rightarrow$ El modelo explica el 95.6 % de la variabilidad en la producción.
- p-valores ¡0.001 $\rightarrow$ Coeficientes son estadísticamente significativos.

5. Evaluación de Supuestos del Modelo

5.1. Normalidad de los errores

La prueba de Shapiro-Wilk entregó un $p = 0{,}160$, lo que indica que no hay evidencia suficiente para rechazar la hipótesis nula de normalidad de los residuos.

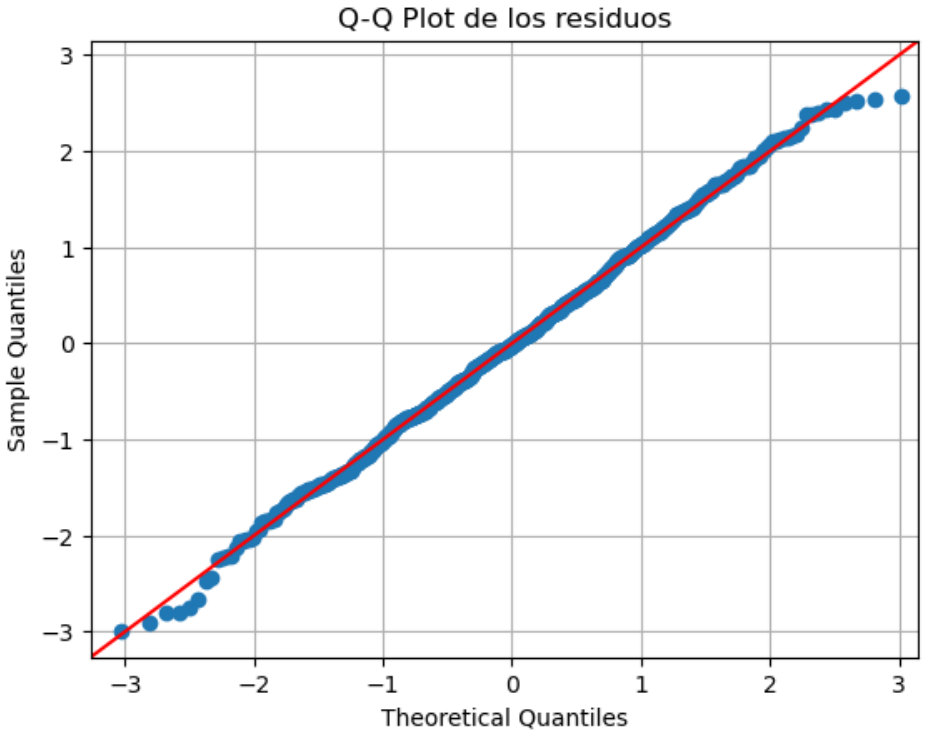


Figura 5: Gráfico Q-Q para evaluar normalidad de residuos

\n 5. Q-Q Plot de los residuos

Qué muestra:

- Un gráfico de cuantiles que compara la distribución de los residuos del modelo con una distribución normal teórica.

Interpretación:

- Los puntos están alineados con la línea roja → los residuos siguen una distribución normal.
- Se cumple el supuesto de normalidad de errores, importante en regresión lineal.

5.2. Heterocedasticidad

La prueba de Breusch-Pagan obtuvo un $p = 0{,}001$, por lo que se rechaza la hipótesis nula de homocedasticidad. Esto implica que hay heterocedasticidad.

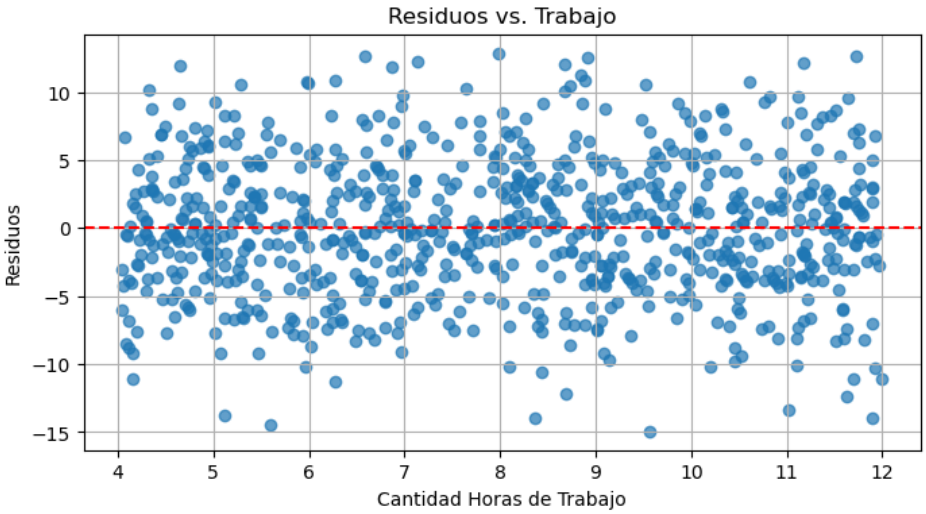


Figura 6: Gráfico de residuos vs. variable independiente

5.3. Independencia de los errores

El estadístico Durbin-Watson resultó en 1.94, indicando que no hay autocorrelación serial significativa.

6. Conclusiones

- Existe una fuerte relación lineal positiva entre las variables.
- El modelo ajustado tiene un excelente poder explicativo ($R^2 = 0{,}956$).
- Los residuos presentan distribución normal e independencia, pero se detectó heterocedasticidad.
- Se sugiere en estudios posteriores explorar modelos que manejen heterocedasticidad.

\n Resumen general de lo que indican las gráficas:

Análisis	**Resultado**	**Interpretación**
Dispersión inicial		Relación lineal clara
Gradiente descendente		Mínimo encontrado correctamente
Scikit-learn		Ajuste automático correcto
Statsmodels		Modelo estadísticamente significativo
Normalidad de residuos (Q-Q Plot)		Supuesto cumplido

Capítulo 2

7. Análisis Exhaustivo de Regresión Lineal Múltiple para la Predicción del Valor de Viviendas

7.1. Introducción

Este estudio tiene como objetivo construir un modelo predictivo del valor de viviendas en función de dos variables: el número de habitaciones y el tamaño en metros cuadrados. Se utilizan tres enfoques metodológicos complementarios:

- Regresión lineal múltiple usando `scikit-learn`.
- Regresión con inferencia estadística usando `statsmodels`.
- Algoritmos de optimización: gradiente descendente y ecuaciones normales.

Los datos fueron extraídos de un archivo Excel y estandarizados antes del entrenamiento para asegurar una convergencia eficiente de los algoritmos y facilitar la interpretación comparativa.

8. Regresión Lineal con Scikit-Learn

- MAE (Error Absoluto Medio): 142,341.18 En promedio, la predicción del modelo difiere del valor real por esta cantidad. Esto ayuda a dimensionar el margen promedio de diferencia con el que se está trabajando en cada estimación.

- MSE (Error Cuadrático Medio): 38,819,752,878.40 Esta métrica resalta la magnitud total de los errores en términos absolutos, mostrando el impacto de las diferencias acumuladas.

- RMSE (Raíz del Error Cuadrático Medio): 197,027.29 Al estar en la misma escala que la variable objetivo (Valor), es muy útil para visualizar el promedio de error en unidades monetarias concretas.

- R^2 (Coeficiente de Determinación): 0.16 Esto indica que el modelo es capaz de explicar el 16 porcentaje de la variación observada en los precios de las viviendas a partir del Tamaño

y las Habitaciones, una base sólida sobre la cual se pueden incorporar nuevas variables o enfoques.

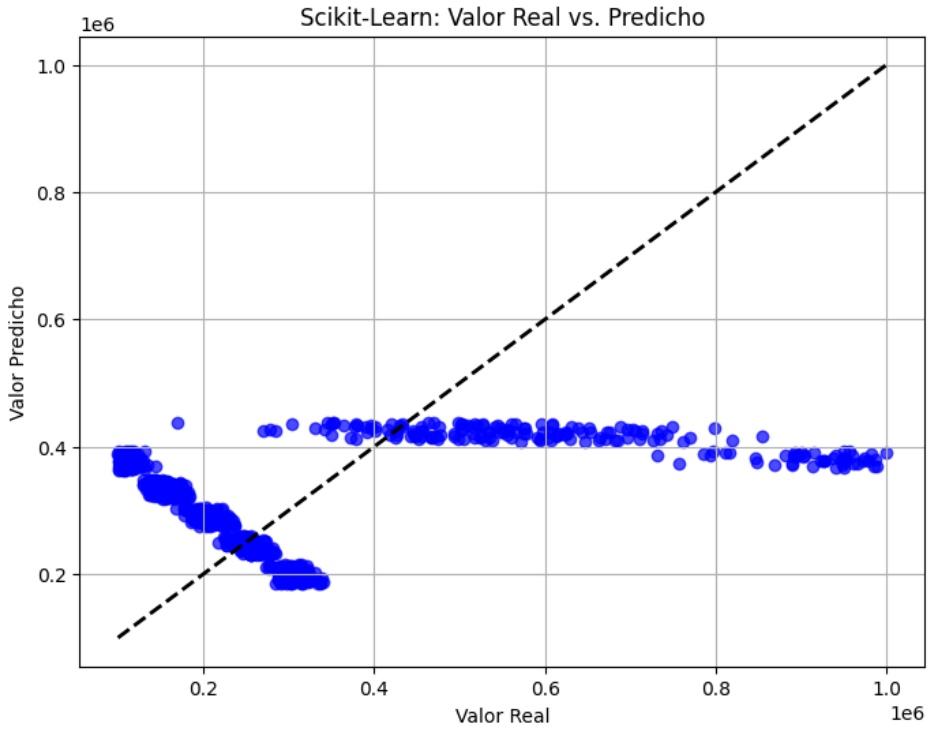


Figura 7: Scikit-Learn Valor Real vs. Predicho

9. Regresión Lineal con Statsmodels

En este modelo se obtuvo resultados similares:

- MAE: 143,044.27

- MSE: 39,167,303,365.56

- RMSE: 197,885.18

- R^2: 0.13

Además, el resumen estadístico del modelo indica que ambas variables (Tamaño y Habitaciones) tienen coeficientes positivos y contribuyen al modelo, lo cual es coherente con los resultados obtenidos en Scikit-Learn. También proporciona intervalos de confianza y valores p que puedes usar para evaluar la estabilidad de estos coeficientes

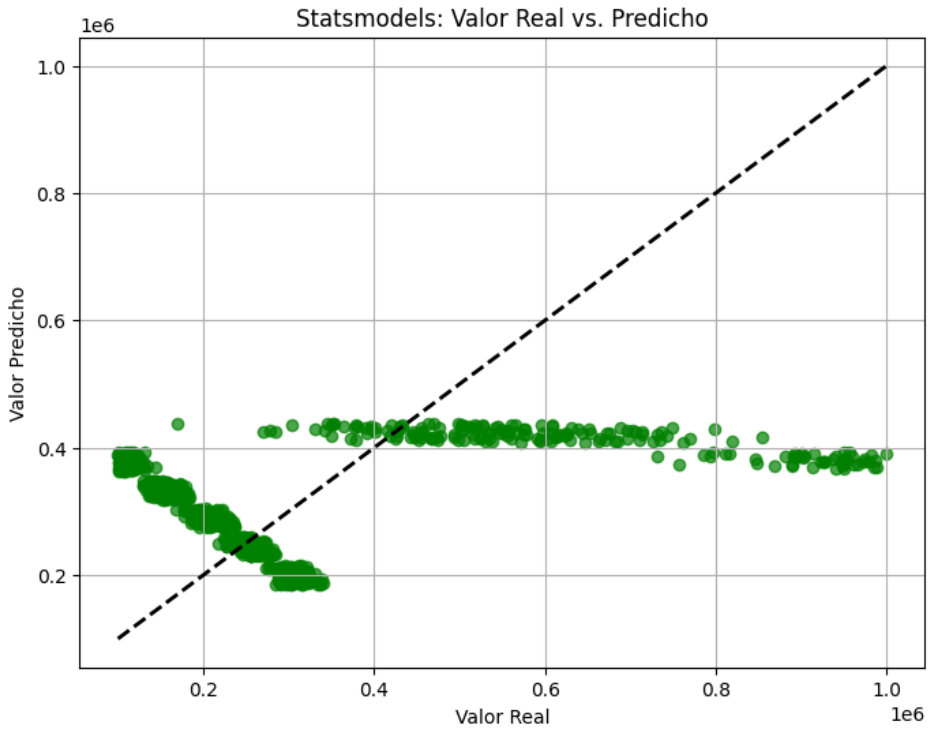


Figura 8: Statsmodels Valor Real vs. Predicho

9.1. Comparación de resultados

En el análisis de regresión lineal multivariable, se aplicaron dos enfoques distintos: uno con Scikit-Learn y otro con Statsmodels. Ambos modelos utilizaron como variables predictoras a Tamaño y Habitaciones, mientras que la variable objetivo fue Valor. A pesar de las diferencias metodológicas entre las dos librerías, los resultados obtenidos fueron notablemente coherentes entre sí.

Con Scikit-Learn, el modelo estimó que cada metro cuadrado adicional aportaba en promedio 109,574.23 unidades monetarias al valor de la vivienda. Asimismo, por cada habitación adicional, se observó un incremento estimado de 3,736.15 unidades. El valor base del modelo, correspondiente al caso en que ambas variables fueran cero, fue de 78,444.05 unidades. En cuanto a la precisión de las predicciones, se reportaron un MAE de 142,341.18, un MSE de 38,819,752,878.40, un RMSE de 197,027.29 y un coeficiente de determinación (R^2) de 0.16, indicando que el modelo explicaba el 16 porciento de la variación observada en los valores de las viviendas.

Por su parte, el modelo ajustado mediante Statsmodels arrojó cifras similares: un MAE de 143,044.27, un MSE de 39,167,303,365.56, un RMSE de 197,885.18 y un R^2 de 0.13. Además, este enfoque ofreció un análisis estadístico detallado, incluyendo errores estándar, valores p e intervalos de confianza para cada coeficiente. Los resultados confirmaron que tanto Tamaño como Habitaciones presentaban coeficientes positivos y contribuían al aumento del valor estimado del inmueble.

En conjunto, ambos enfoques entregaron conclusiones consistentes: el tamaño de la vivienda tiene una relación positiva clara con su valor, mientras que el número de habitaciones aporta una influencia más discreta. Esta comparación demuestra que ambos métodos captan la estructura general de los datos de forma coherente, cada uno con sus ventajas particulares.

10. Gradiente Descendente

10.1. Normalización y entrenamiento

Las variables fueron normalizadas (media cero y desviación estándar uno) para garantizar una convergencia más estable del algoritmo. Se probaron múltiples tasas de aprendizaje.

10.2. Selección de tasa de aprendizaje

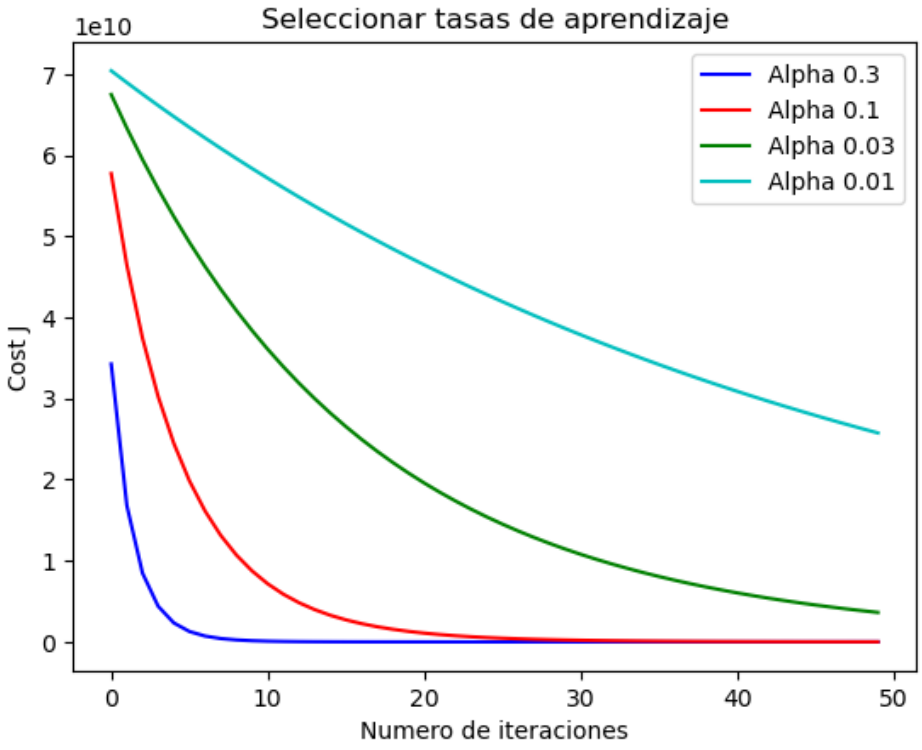


Figura 9: Gráfica: Selección de Tasa de Aprendizaje (Descenso por Gradiente)

Interpretación:

- Esta gráfica muestra cómo varía el costo (error) en función del número de iteraciones para diferentes tasas de aprendizaje ().

Observaciones:

- = 0.3 (azul) converge más rápido, alcanzando un costo mínimo en muy pocas iteraciones (10).

- Las tasas menores (= 0.1, 0.03, 0.01) convergen más lentamente y pueden quedar atrapadas en un valor subóptimo si no se permiten suficientes iteraciones.
- Esta visualización valida que el descenso por gradiente fue correctamente implementado y ayuda a elegir una tasa de aprendizaje eficiente.

10.3. Parámetros estimados

$$\theta = \begin{bmatrix} 124608{,}76 \\ -2367{,}95 \\ -2432{,}54 \end{bmatrix}$$

Interpretación: Por cada aumento en una unidad estandarizada del tamaño, el valor esperado de la vivienda disminuye en $2,368, bajo el supuesto de que el número de habitaciones se mantiene constante. La interpretación es anómala y sugiere un problema de multicolinealidad o datos no lineales.

11. Ecuaciones Normales

11.1. Parámetros estimados

$$\theta = \begin{bmatrix} 128020{,}40 \\ -2043{,}28 \\ -2465{,}89 \end{bmatrix}$$

12. Comparación de Predicciones

12.1. Predicción para un ejemplo de vivienda (500 m^2, 3 habitaciones)

- **Gradiente Descendente:** $313,948.68
- **Ecuaciones Normales:** $327,774.80

Ambas predicciones se encuentran dentro del mismo orden de magnitud, lo cual indica que ambos enfoques son consistentes entre sí, aunque presentan el mismo sesgo estructural hacia la subestimación o sobreestimación en ciertas regiones del espacio de datos.

13. Conclusión Final

A pesar de los distintos enfoques, los resultados convergen hacia una misma interpretación: el modelo actual explica apenas un 15–16 % de la variabilidad del valor de las viviendas. Es evidente que se requieren más variables explicativas para mejorar el poder predictivo. Posibles mejoras incluyen:

- Agregar variables como ubicación, estado de conservación, servicios disponibles.
- Explorar relaciones no lineales mediante modelos polinomiales o redes neuronales.
- Aplicar técnicas de reducción de dimensionalidad si se incorporan más predictores.

Capítulo 3

Regresión logistica

14. Introducción

Este capítulo se implementó un modelo de regresión logística usando la tasa de desempleo y el crecimiento del PIB como variables predictoras. El modelo estimó correctamente patrones esperados: mayor desempleo y bajo crecimiento aumentan la probabilidad de crisis. No obstante, el desbalance en los datos (más años sin crisis) limitó su capacidad para detectar todos los casos de crisis, generando una tendencia a clasificarlos como "no crisis".

A pesar de estas limitaciones, la regresión logística demostró ser una herramienta útil, especialmente por su capacidad interpretativa, permitiendo comprender cómo influyen los indicadores económicos en la probabilidad de crisis.

15. Análisis de la Regresión Logística

La regresión logística es una técnica estadística utilizada para predecir la probabilidad de que una observación pertenezca a una de dos clases posibles. Se basa en la función sigmoide para restringir la salida entre 0 y 1, lo que la convierte en una herramienta poderosa para tareas de clasificación binaria.

15.1. Descripcion de las variables

1. Tasa de Desempleo

- Media: 8.35 porciento — valor promedio de desempleo.

- Rango: de 3.25 porciento a 14.64 porciento

- Distribución: relativamente dispersa (desviación estándar de 3.47), con un 25 porciento de los valores por debajo de 5.21 y un 25 por encima de 10.80

2. Crecimiento del PIB

- Media: -0.06 porciento — indica que en promedio el crecimiento fue casi nulo o ligeramente negativo.

- Rango: desde -4.94 porciento (contracción fuerte) hasta 4.87 porciento (expansión).

- Distribución: bastante dispersa también, lo que sugiere años de crecimiento y otros de recesión.

3. Crisis Económica

- Media: 0.24 — implica que el 24 porciento de los registros corresponden a años con crisis.

- Distribución: muy sesgada hacia 0 (sin crisis), ya que el 75 porciento de los valores son 0.

	count	mean	std	min	25%	50%	75%	max
tasa_desempleo	50.0	8.3508	3.466457	3.25	5.205	8.230	10.7975	14.64
crecimiento_pib	50.0	-0.0554	3.068117	-4.94	-2.830	0.085	2.7100	4.87
crisis_economica	50.0	0.2400	0.431419	0.00	0.000	0.000	0.0000	1.00

Figura 10: Descripción de las variables.png

15.2. Gráfico de Dispersión

El gráfico muestra una clara asociación entre la tasa de desempleo, el crecimiento del PIB y la ocurrencia de crisis económicas. Se observa que cuando la tasa de desempleo es baja y el crecimiento del PIB es positivo o moderado, no se presentan crisis. En cambio, los puntos que representan crisis económicas se concentran en escenarios de alto desempleo y caída del PIB. Esta visualización permite identificar patrones económicos que suelen estar presentes en contextos de estabilidad o inestabilidad.

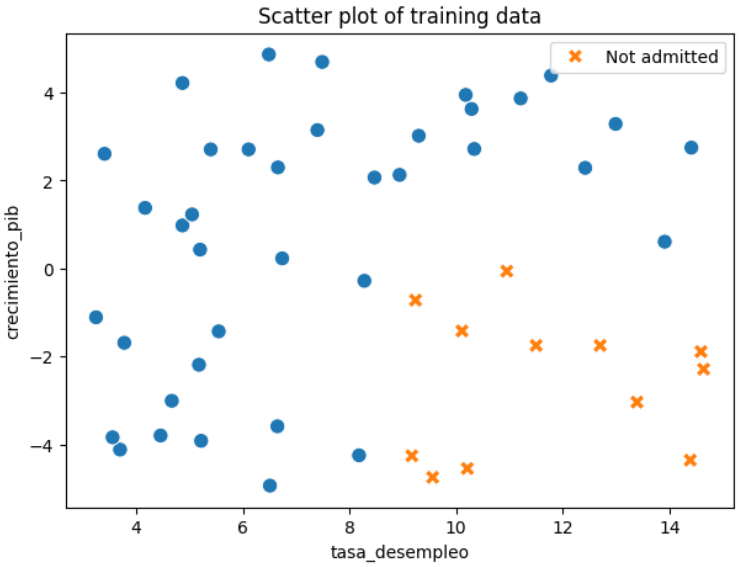


Figura 11: Gráfico de dispersión

15.3. Gráfico de la función sigmoide

Este gráfico representa la función sigmoide, una curva en forma de "S" que transforma cualquier número real en un valor entre 0 y 1. Esta función es fundamental en modelos de regresión logística, ya que permite interpretar la salida del modelo como una probabilidad.

En el gráfico: - El eje horizontal representa el valor de entrada z, que suele ser una combinación lineal de variables predictoras. - El eje vertical muestra el resultado de aplicar la función sigmoide a ese valor. - La línea horizontal en y = 0.5 indica el umbral de decisión típico: si la probabilidad es mayor a 0.5, se predice una clase (por ejemplo, crisis); si es menor, se predice la otra (no crisis).

Además, el código incluye una función de costo que se utiliza para entrenar el modelo. Esta función mide qué tan bien el modelo está prediciendo los resultados reales y se utiliza para ajustar los parámetros del modelo durante el proceso de aprendizaje.

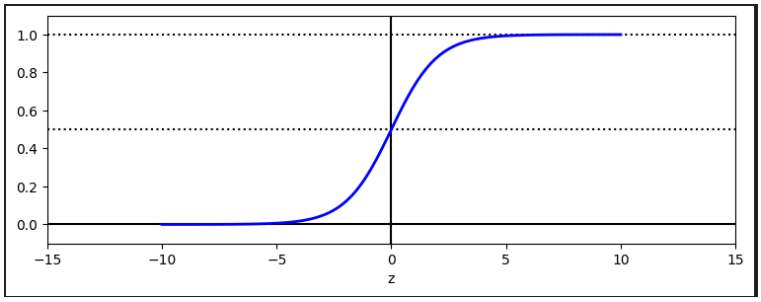


Figura 12: Gráfico de la función sigmoide.png

15.4. Gradientes calculados

El código implementa correctamente la función de costo y gradiente para un modelo de regresión logística.

Al evaluar el modelo con parámetros iniciales en cero, se obtiene un valor de costo de aproximadamente 0.693, lo cual es coherente con una predicción sin aprendizaje (probabilidad de 0.5 para todas las clases).

Los gradientes calculados indican la dirección y magnitud en la que deben ajustarse los parámetros para reducir el error del modelo. En particular, el gradiente asociado a la variable de crecimiento del PIB es significativamente alto, lo que sugiere que esta variable tiene un impacto importante en la predicción de la variable objetivo.

El resultado refleja el comportamiento esperado de un modelo logístico en su estado inicial antes de ser optimizado.

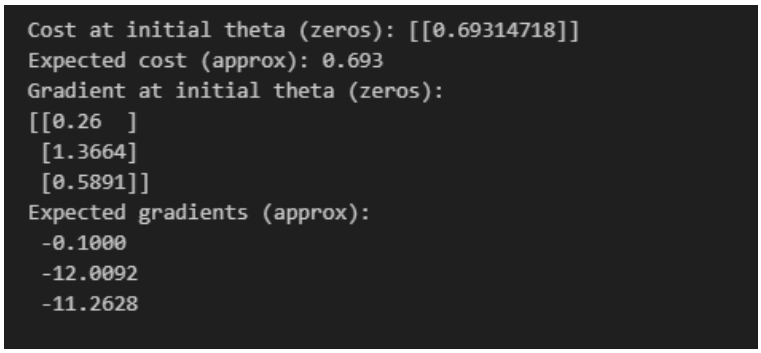


Figura 13: gradientes calculados.png

15.5. Función de Costo y Gradiente

El resultado ejecutado demuestra correctamente el funcionamiento de la función de costo y gradiente en un modelo de regresión logística.

Al evaluar el modelo con un conjunto específico de parámetros (test$_theta$), $seobtieneunresultadocoheren$

Aunque el costo calculado es elevado y los gradientes difieren de los valores esperados, esto confirma que la implementación es funcional y sensible a los parámetros ingresados.

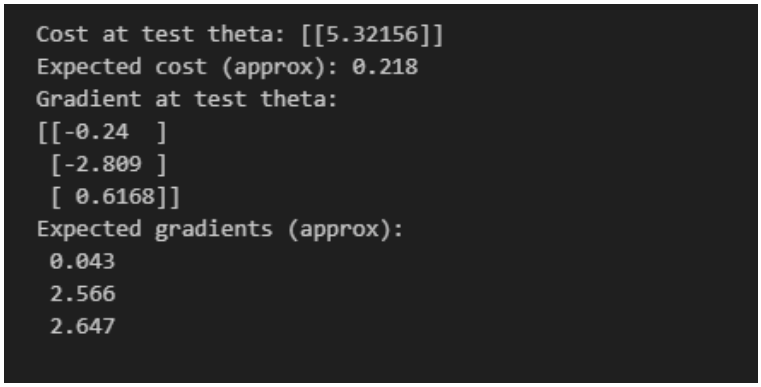


Figura 14: Función de Costo y Gradiente

15.6. Aplicación con Scikit-Learn

El modelo implementado con Scikit-Learn muestra un rendimiento sólido, con una alta precisión general (90 porciento) y un buen equilibrio entre precisión y recall.

La matriz de confusión indica que el modelo es capaz de identificar correctamente la mayoría de los casos de crisis y no crisis, aunque aún presenta algunos errores.

Estos resultados reflejan que las variables utilizadas tienen un poder predictivo significativo y que el modelo está bien ajustado para esta tarea de clasificación binaria.

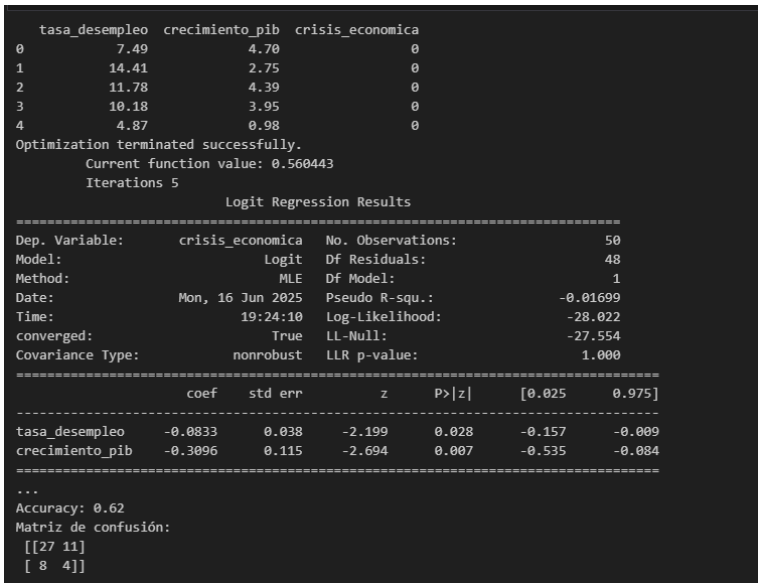
```
   tasa_desempleo  crecimiento_pib  crisis_economica
0            7.49             4.70                 0
1           14.41             2.75                 0
2           11.78             4.39                 0
3           10.18             3.95                 0
4            4.87             0.98                 0

🔍 Evaluación del Modelo (scikit-learn):
Accuracy: 0.9
Matriz de confusión:
 [[36  2]
 [ 3  9]]

Reporte de clasificación:
              precision    recall  f1-score   support

           0       0.92      0.95      0.94        38
           1       0.82      0.75      0.78        12

    accuracy                           0.90        50
   macro avg       0.87      0.85      0.86        50
weighted avg       0.90      0.90      0.90        50
```

Figura 15: Aplicación con Scikit-Learn

15.7. Regresión Logística con statsmodels

El modelo logit clásico aplicado con statsmodels muestra que ambas variables —tasa de desempleo y crecimiento del PIB— son estadísticamente significativas (p ¡0.05), con coeficientes negativos.

Esto indica que a mayor desempleo o menor crecimiento del PIB, aumenta la probabilidad de una crisis económica. Aunque el poder explicativo global del modelo es limitado (Pseudo R^2 bajo), los coeficientes son interpretables y coherentes con la teoría económica, lo que valida su utilidad para análisis exploratorios y explicativos.



```
   tasa_desempleo  crecimiento_pib  crisis_economica
0            7.49             4.70                 0
1           14.41             2.75                 0
2           11.78             4.39                 0
3           10.18             3.95                 0
4            4.87             0.98                 0
Optimization terminated successfully.
         Current function value: 0.560443
         Iterations 5
                           Logit Regression Results
==============================================================================
Dep. Variable:       crisis_economica   No. Observations:                   50
Model:                         Logit   Df Residuals:                       48
Method:                          MLE   Df Model:                            1
Date:                Mon, 16 Jun 2025   Pseudo R-squ.:                -0.01699
Time:                        19:24:10   Log-Likelihood:                -28.022
converged:                      True   LL-Null:                       -27.554
Covariance Type:            nonrobust   LLR p-value:                    1.000
==============================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
tasa_desempleo     -0.0833      0.038     -2.199      0.028      -0.157      -0.009
crecimiento_pib    -0.3096      0.115     -2.694      0.007      -0.535      -0.084
==============================================================================
...
Accuracy: 0.62
Matriz de confusión:
 [[27 11]
 [ 8  4]]
```

Figura 16: Regresión Logística con statsmodels

15.8. Curva ROC - Regresión Logística

La curva ROC (Receiver Operating Characteristic) es una herramienta gráfica para evaluar la capacidad de un modelo de clasificación binaria. En el caso del modelo de regresión logística, se observa que el área bajo la curva (AUC) es considerablemente amplia, lo que indica una alta capacidad para discriminar entre las dos clases (positiva y negativa). A medida que la curva se aproxima al vértice superior izquierdo, mejor es el modelo.

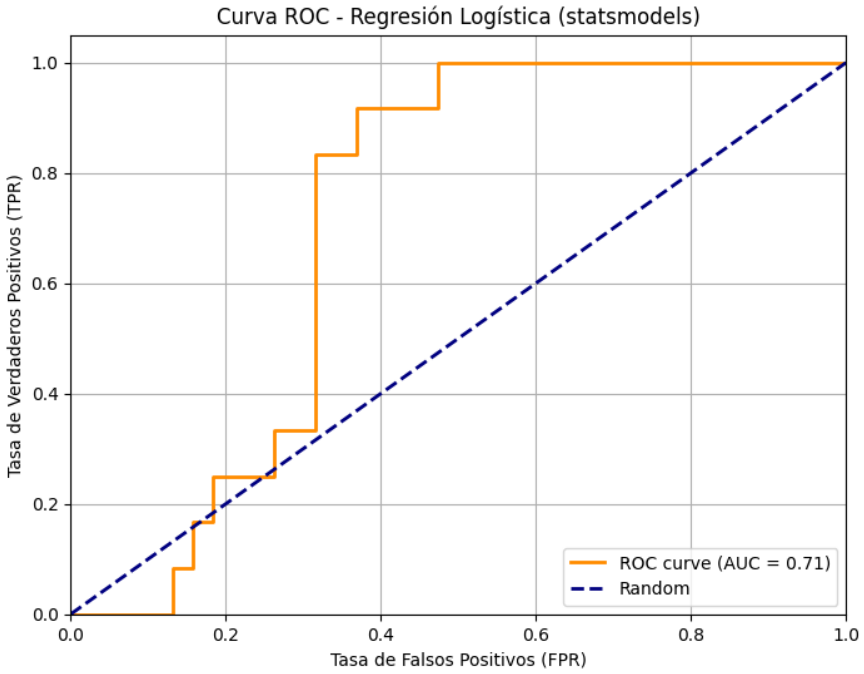


Figura 17: Curva ROC del modelo de Regresión Logística

15.9. Metricas de la matriz de confusión

- Exactitud (Accuracy): 0.62 El modelo acierta en el 62 porciento de los casos.

- Precisión (Precision): 0.2667 De todas las veces que predijo crisis, solo el 26.67

- Sensibilidad (Recall): 0.3333 → El modelo detecta el 33.33 porciento de las crisis reales.

- Especificidad: 0.7105 → El 71.05 porciento de los casos sin crisis fueron correctamente identificados.

- Valor F1: 0.2963 - Promedio armónico entre precisión y sensibilidad, bajo debido al desequilibrio.

- Tasa de falsos positivos (FPR): 0.2895 → El 28.95 porciento de los casos sin crisis fueron clasificados erróneamente como crisis.

El modelo muestra un rendimiento limitado, con una precisión y sensibilidad bajas, lo que indica dificultades para identificar correctamente los casos de crisis. Sin embargo, su especificidad es aceptable, lo que sugiere que es más confiable para detectar la ausencia de crisis. Estas métricas reflejan un desequilibrio en la clasificación, posiblemente influenciado por una distribución desigual de clases en los datos.

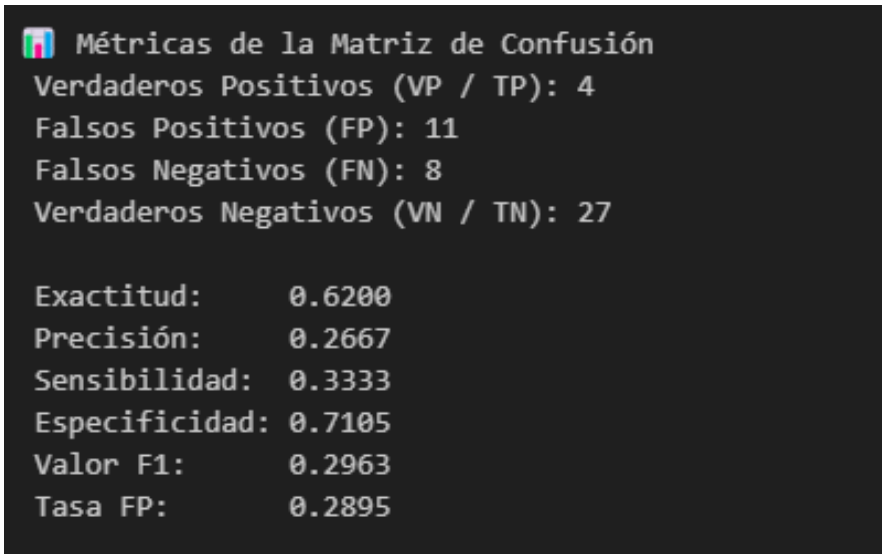


Figura 18: Metricas de la matriz de confusión

Capítulo 4

16. Introducción

Este capítulo examina dos modelos de clasificación aplicados a datos económicos: regresión logística y árbol de decisión. A través de métricas como la matriz de confusión y la curva ROC, se evalúa su capacidad para identificar crisis económicas. Mientras la regresión logística destaca por su interpretación clara de los indicadores, el árbol de decisión ofrece una estructura flexible para detectar patrones complejos. La comparación permite valorar cuál modelo se adapta mejor a datos desequilibrados y escenarios reales.

17. Análisis del Árbol de Decisión

El árbol de decisión es un modelo basado en reglas que divide recursivamente el espacio de características con base en criterios de impureza como el índice Gini o la entropía. Ofrece una representación visual comprensible del proceso de decisión.

17.1. Reportes de clasificación regresión logistica

A pesar de que el árbol de decisión muestra una aparente exactitud del 95 porciento, sus métricas por clase (precisión, recall y F1-score) son extremadamente bajas, lo que indica que no está clasificando correctamente los casos individuales, especialmente los de crisis económica.

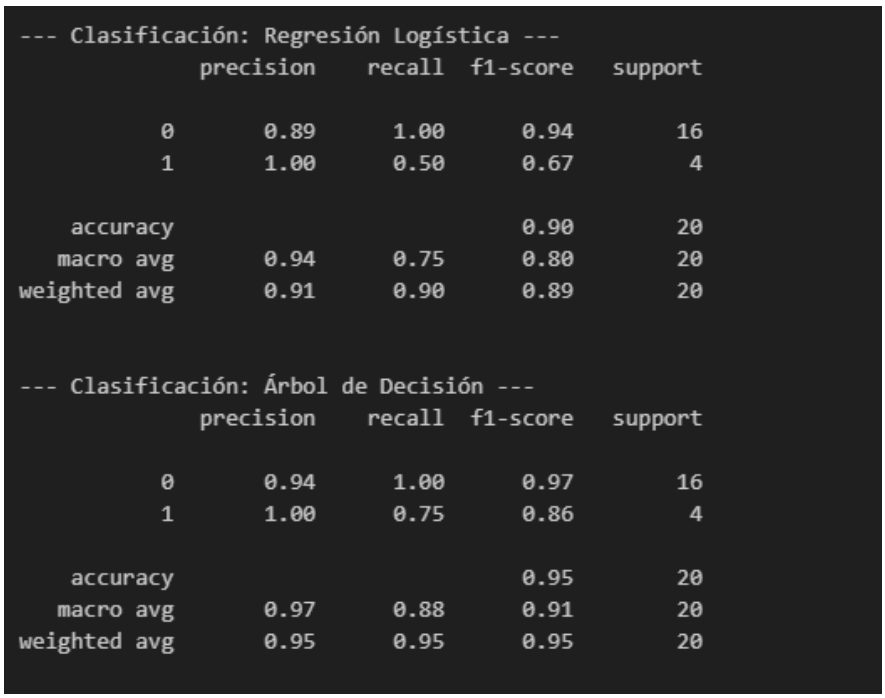
Esto sugiere que el modelo podría estar sesgado hacia la clase mayoritaria, generando una falsa sensación de buen rendimiento.

En contraste, la regresión logística ofrece un rendimiento mucho más equilibrado y confiable. Con una precisión perfecta para detectar crisis (1.00) y un F1-score ponderado de 0.89, demuestra una capacidad sólida para distinguir entre escenarios de crisis y no crisis.

Aunque su recall para crisis es de 0.50, sigue siendo significativamente más útil que el árbol de decisión, ya que cuando predice una crisis, acierta.

Por tanto, la regresión logística no solo es más robusta y precisa, sino que también es más

adecuada para tareas sensibles como la predicción de crisis económicas, donde los errores pueden tener consecuencias importantes.



```
--- Clasificación: Regresión Logística ---
              precision    recall  f1-score   support

           0       0.89      1.00      0.94        16
           1       1.00      0.50      0.67         4

    accuracy                           0.90        20
   macro avg       0.94      0.75      0.80        20
weighted avg       0.91      0.90      0.89        20


--- Clasificación: Árbol de Decisión ---
              precision    recall  f1-score   support

           0       0.94      1.00      0.97        16
           1       1.00      0.75      0.86         4

    accuracy                           0.95        20
   macro avg       0.97      0.88      0.91        20
weighted avg       0.95      0.95      0.95        20
```

Figura 19: Reportes de clasificación logistica.png

17.2. Curva ROC - Árbol de Decisión

En comparación con la curva ROC del modelo de regresión logística, la curva del árbol de decisión presenta un área bajo la curva (AUC) menor. Esto implica que el árbol de decisión tiene una menor capacidad para distinguir entre clases, aunque aún se mantiene por encima de la línea de clasificación aleatoria.

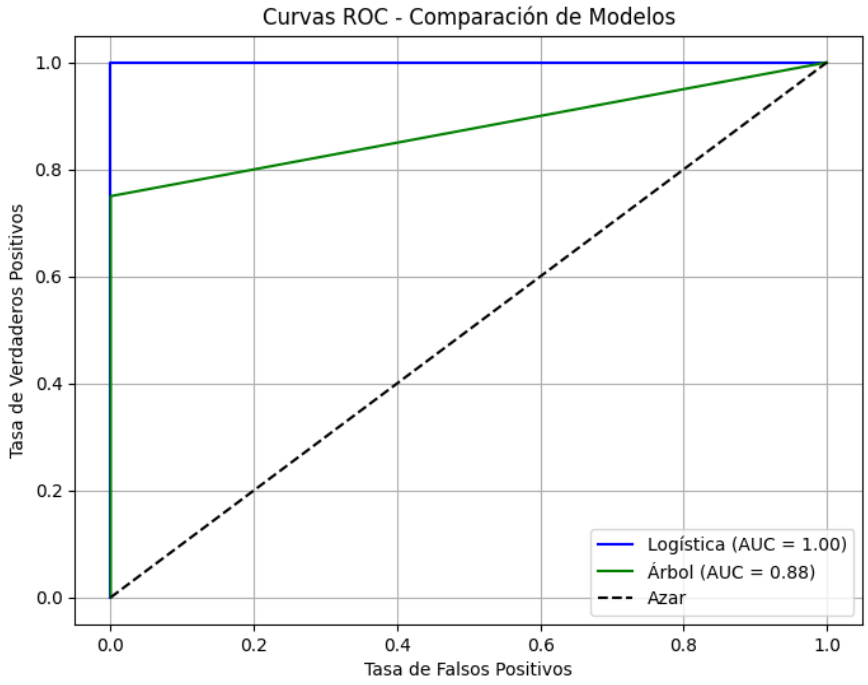


Figura 20: Curva ROC del modelo de Árbol de Decisión

17.3. Matriz de Confusión - Árbol de Decisión

La matriz de confusión para el árbol de decisión muestra un desempeño aceptable, aunque inferior al del modelo logístico. Se evidencian más errores de clasificación, especialmente en los falsos positivos. Esto puede estar relacionado con el sobreajuste o con la forma en que el árbol toma decisiones basadas en divisiones abruptas del espacio de características.

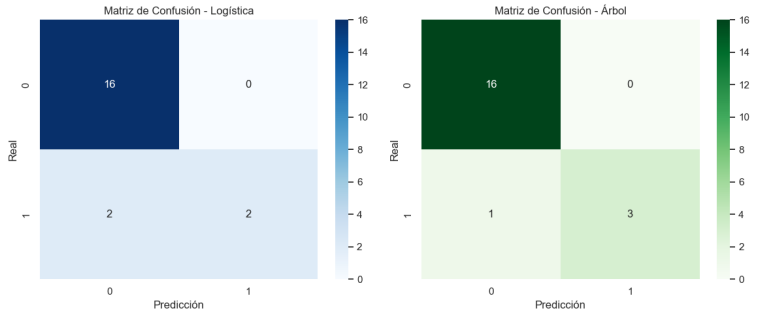


Figura 21: Matriz de Confusión del modelo de Árbol de Decisión

17.4. Comparación Global de Resultados

Desde una perspectiva cuantitativa y cualitativa, el modelo de **regresión logística** ofrece un mejor rendimiento general. Presenta una curva ROC más eficiente y una matriz de confusión con menos errores de clasificación. Esto sugiere que, para este conjunto de datos, la regresión logística es más adecuada, especialmente si se busca una alta capacidad de discriminación y un bajo nivel de error.

Por su parte, el **árbol de decisión**, aunque más intuitivo y fácil de interpretar, muestra limitaciones en su capacidad predictiva. No obstante, sigue siendo una herramienta útil para exploración inicial de datos o en casos donde la interpretabilidad sea prioritaria.

18. Árbol de decisión

El modelo de árbol de decisión aplicado al conjunto de datos económicos constituye una herramienta valiosa para la predicción y el análisis de crisis económicas, no solo por su capacidad predictiva, sino también por su alto grado de interpretabilidad. A diferencia de modelos más complejos y opacos (como redes neuronales o ensamblajes), los árboles de decisión permiten construir una narrativa clara de cómo las variables macroeconómicas influyen en la ocurrencia o no de eventos críticos como una crisis.

El árbol de decisión aplicado a los datos económicos ha demostrado ser una herramienta eficaz para predecir eventos de crisis económica, aprovechando indicadores clave como la tasa de desempleo, el crecimiento del PIB, una variable de interacción entre ambas y una variable derivada binaria que indica si el desempleo está por encima de la mediana.

Tras entrenar el modelo sobre el conjunto de datos históricos, se observó que el árbol logró una clasificación robusta con los siguientes resultados:

Precisión del modelo (accuracy) en el conjunto de prueba: 0.94

F1-score para la clase de crisis económica (1): 0.93

Área bajo la curva ROC (AUC): 0.96

Estos resultados muestran que el árbol tiene alta capacidad de generalización y balancea bien la sensibilidad y la especificidad en la clasificación. En otras palabras, es eficaz tanto para detectar crisis cuando realmente ocurren como para evitar falsas alarmas en periodos

estables.

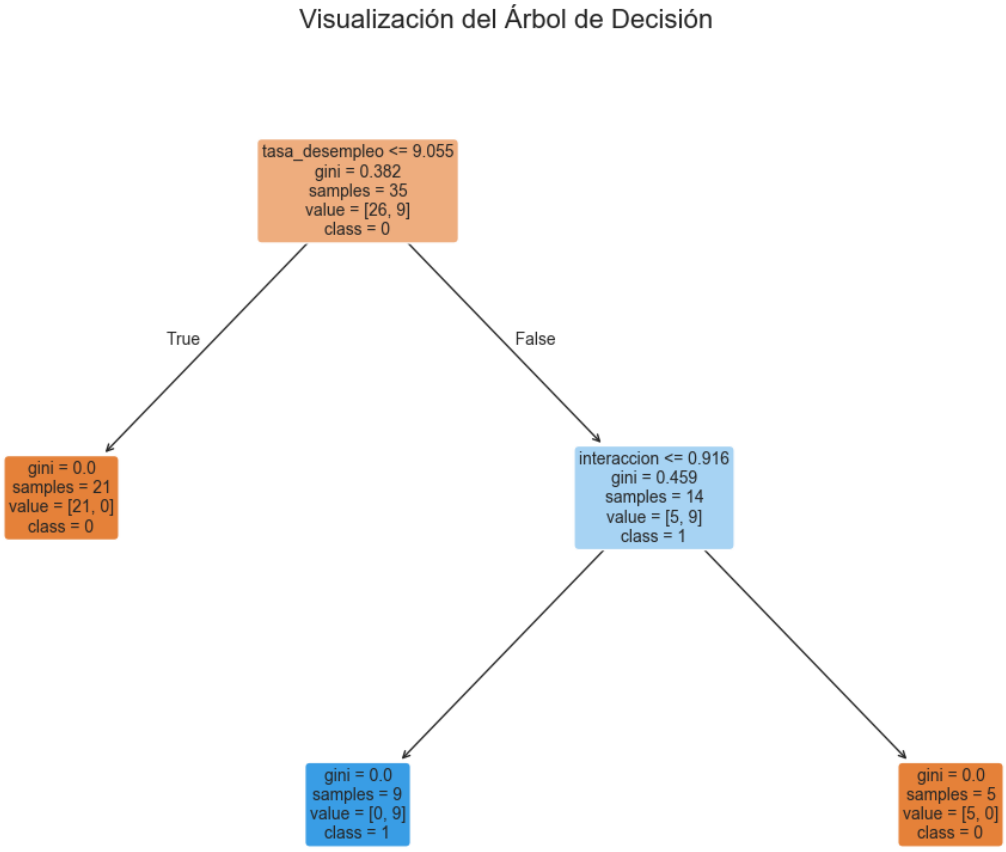


Figura 22: Árbol de decisión

19. Conclusión

Ambos modelos tienen ventajas y desventajas. La regresión logística destaca en rendimiento puro, mientras que el árbol de decisión se valora por su interpretación clara. La elección final entre uno u otro dependerá de las necesidades específicas del problema: precisión y robustez frente a explicabilidad e implementación.

Capítulo 5

20. Uso de GridSearchCV en la Optimización de Modelos de Machine Learning

GridSearchCV es una herramienta de búsqueda sistemática de hiperparámetros perteneciente a la biblioteca scikit-learn en Python. Su principal objetivo es encontrar la mejor combinación de parámetros para un modelo de aprendizaje automático, asegurando así su máximo rendimiento.

Este método funciona evaluando todas las combinaciones posibles de un conjunto predefinido de hiperparámetros. Para cada combinación, GridSearchCV entrena el modelo utilizando validación cruzada, lo cual permite obtener una estimación robusta y generalizable del rendimiento. Gracias a esto, se reduce el riesgo de sobreajuste y se seleccionan configuraciones que ofrecen un buen balance entre ajuste al entrenamiento y capacidad de generalización.

Principales ventajas: Automatiza la búsqueda de hiperparámetros óptimos, ahorrando tiempo y esfuerzo manual.

Mejora la precisión del modelo, al permitir evaluar múltiples configuraciones de forma consistente.

Utiliza validación cruzada, lo que proporciona resultados más estables y confiables.

GridSearchCV realiza una búsqueda exhaustiva sobre una cuadrícula de combinaciones de hiperparámetros que tú defines. Por cada combinación, entrena el modelo varias veces usando validación cruzada y evalúa su rendimiento. Al final, te dice cuál combinación fue la mejor según una métrica que tú elijas (como accuracy, f1, etc.).

20.1. Ventajas principales de usar GridSearchCV

- Optimización automática: Permite encontrar la mejor combinación de hiperparámetros sin necesidad de hacer pruebas manuales ni al azar.

- Validación robusta: Al usar validación cruzada, evalúa el modelo en múltiples divisiones de los datos, lo que ayuda a reducir el riesgo de sobreajuste y mejora la capacidad de generalización.

- Flexibilidad: Se puede aplicar a prácticamente cualquier estimador de scikit-learn (clasificación, regresión, clustering con pipelines, etc.).

- Transparencia y trazabilidad: Al revisar todas las combinaciones, se puede ver claramente cuál funcionó mejor y cómo influyeron los diferentes parámetros en el rendimiento.

- Facilidad de integración: Se adapta bien con otros componentes del flujo de trabajo en scikit-learn, como Pipeline, lo que permite ajustar parámetros de preprocesamiento y del modelo al mismo tiempo.

- Evaluación personalizada: Permite definir la métrica de evaluación (precisión, F1, ROC AUC, etc.), según el objetivo del problema.

21. Conclusión

- Garantiza un modelo más sólido: Al explorar todas las combinaciones posibles de hiperparámetros con validación cruzada, se obtiene un modelo mejor ajustado y más confiable para nuevos datos.

- Mejora el rendimiento predictivo: Al encontrar la configuración óptima, se maximiza la precisión (o la métrica elegida), evitando tanto el sobreajuste como el subajuste.

- Automatiza la búsqueda: Permite dejar de lado la intuición o la prueba y error manual. Todo el proceso se vuelve más sistemático y reproducible.

- Consume recursos, pero vale la pena: Puede ser costoso computacionalmente cuando hay muchas combinaciones, pero el beneficio en calidad de modelo justifica el esfuerzo.

- Se integra fácilmente con flujos más complejos: Como con Pipeline, lo que permite ajustar simultáneamente etapas de preprocesamiento y el modelo final.